# 線型回帰モデルにおける異常値の検出と頑健推定におよぼす独立変数観測値の影響

原田桂一郎

目　　次



## 1　はじめに

経済現象を計量経済学的に分析するためには，経済現象を経済変数間の因果関係として捉え，これを明示的に表現したモデル，そして統計データを必要とする。モデルは，現象を構成する本質的な要因を変数とし，理論に基づいて変数の一次式によって現象を表現したものである。統計データは無数の要因から成る現象を観測することによって与えられるものである。

モデルは多くの要因を変数として組み込んで複雑化すれば，現実（データ）と合致したものとすることができる。つまり，理論とデータが整合的であるようなモデルをいくらでも構築することが可能なのである。したがって，計量モデルを用いて理論的命題（仮説）を経験（観測結果）によっ

て検証することは不可能となる。

計量経済分析において用いられるモデルは，理論を統計データによって検証する，あるいは理論が統計データと整合的であるか否かを調べるためのものではない。経済理論を正しいものと考え，それに基づいて経済関係を計量モデル化する。そして統計データによって推定した経済関係（構造）によって経済現象の数量分析や予測が行われる。経済理論を統計データよりも重視し優先させる。

ここで，現象の観測によって理論が検証できないとすると，経済理論の有意性の規準はこどに求められるか，という重要な問題が存在するが，この点については本稿の主題から離れたものとなるので取り扱わないことにしたい。

さて，本稿で取り扱う計量モデル（単一方程式モデルを考える）は次の型式となる。すなわち，経済変数間の関係を一次式をもって表わした理論モデルは現実の経済現象の近似表現であるから，現実と理論モデルの乖離は誤差項で埋め合はされ線型回帰モデルの型式となる。誤差項は無数の独立な要因（誤差）の和であるとして，中心極限定理により近似的に正規分布にしたがう確率変数と仮定される。正規分布は裾野が薄い分布であるから，大きな誤差，あるいは経済関係を大きく左右するような攪乱要因が発生する確率が低いことを仮定している。さらに，データ観測期間を通じて構造変化[1]が生じないことを仮定すると，最小二乗推定される回帰係数（経済関係）は極めて安定的なものになる。

しかし，データ観測期間中に構造変化がない場合でも，通常でない事態の発生に関連し攪乱要因を含んだ（経済変数の）観測値，および様々な原因による誤差を含んだ観測値の存在は，それ等を推定に使用するか否かによって結果が大いに異なることが，実証分析の経験から明らかとなっている。こうした観測値は，従属変数の観測値の中に異常値（outliers）となって出現し，安定した経済関係の推定を妨げる。

このような攪乱要因や誤差が変数から分離できたものとして（現実には

不可能), 回帰モデルの誤差項において処理すると次のようなものとなる。その第一は，少数であるが低い確率で発生する作用の大きな攪乱要因を考慮する，第二は誤差項を裾野の厚い分布（正規分布ではなく，大きな誤差あるいは攪乱要因が生じる確率が低くない）を仮定する，そして第三は両者を合併したものとする，が考えられる。いずれのケースで処理しても，回帰係数の最小二乗推定量の効率は著しく低下する。

異常値が存在する場合の現実的対応は，一定の基準により異常値を検出・除去し，正規母集団から抽出した標本観測値のみを用いて回帰係数の最小二乗推定を行うことである。また，異常値の検出・除去をしないが，最小二乗推定とは別の方式ですすめられる頑健推定（robust estimation）が開発されており，この方法を採用することもできる。

ところが，独立変数の観測値の中に，他の観測値に比べて特に大きな値を示すもの（high leverage）が存在すると，これが異常値の検出と回帰係数の頑健推定に著しい影響を与える。経済関係を線型回帰モデルによって分析する場合，独立変数観測値の high leverage は，究極的に安定した経済関係（回帰係数）を推定することを妨げる。

以下において，線型回帰モデルの係数の最小二乗推定に与える観測値の影響，独立変数観測値の high leverage の検出，異常値の検出と頑健推定における high leverage の影響，について考察する。こらを通して，観測値から異常値や high leverage を検出・除去して，安定した回帰係数を推定する方途を示す。

## 注

① 構造変化の問題については，Harada, K. and R. Shirogane : "Recursive Estimation Analysis for Examining the Stability of Regression Relationships," *Memoirs of the Kokushikan University Computer Center*, No. 6 (1985), 1—20.

## 2 回帰係数の最小二乗推定に与える観測値の影響

次の標準線型回帰モデルを考える。

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}, \tag{2-1}$$

$\boldsymbol{y}$：従属（被説明）変数ベクトル $\boldsymbol{y}=(y_1\cdots y_n)'$,

$\boldsymbol{X}$：独立（説明）変数行列 $(n\times p)$, $\mathrm{rank}(\boldsymbol{X})=p$, (2-2)

$\boldsymbol{\beta}$：回帰係数ベクトル $\boldsymbol{\beta}=(\beta_1\cdots\beta_p)'$,

$\boldsymbol{u}$：誤差ベクトル $\boldsymbol{u}=(u_1\cdots u_n)'$,

$E(\boldsymbol{u})=0$, $V(\boldsymbol{u})=\sigma^2\boldsymbol{I}$, $\boldsymbol{I}(n\times n)$,

そして，(2-1) が推定されたものを，

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e}, \tag{2-3}$$

$\boldsymbol{y}$：(2-1) に同じ，

$\boldsymbol{X}$：(2-1) に同じ，

$\boldsymbol{b}$：$\boldsymbol{\beta}$ の最小二乗推定量，

$$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{2-4}$$

$\boldsymbol{e}$：残差ベクトル $\boldsymbol{e}=(e_1\cdots\cdots e_n)'$

$S^2$：誤差分散 $\sigma^2$ の推定量，

$$S^2=\boldsymbol{e}'\boldsymbol{e}/(n-p), \tag{2-5}$$

と表わす。さらに次の表記を与えておく。

$\boldsymbol{x}_i$：$\boldsymbol{X}$ 行列の第 $i$ 行（ベクトル）

$\boldsymbol{X}=[\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n]'$, $\boldsymbol{X}'=[\boldsymbol{x}_1'\boldsymbol{x}_2'\cdots\boldsymbol{x}_n']$,

$\boldsymbol{X}'\boldsymbol{X}=\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i$

$\boldsymbol{X}(i)=\boldsymbol{X}$行列から第 $i$ 行を除いた行列，

$\boldsymbol{y}(i)$：ベクトルから第 $i$ 要素を除いたベクトル，

$\boldsymbol{b}(i)$：$\boldsymbol{X}$ と $\boldsymbol{y}$ から第 $i$ 行，第 $i$ 要素を除いた場合の $\boldsymbol{\beta}$ の最小二乗推定量，

$$\boldsymbol{b}(i)=[\boldsymbol{X}'(i)\boldsymbol{X}(i)]^{-1}\boldsymbol{X}'(i)\boldsymbol{y}(i), \tag{2-6}$$

$S^2(i)$：$\boldsymbol{X}$ と $\boldsymbol{y}$ から第 $i$ 行，第 $i$ 要素を除いた場合の $\sigma^2$ の推定量，

$$S^2(i)=[\boldsymbol{y}(i)-\boldsymbol{X}(i)\boldsymbol{b}(i)]'[\boldsymbol{y}(i)-\boldsymbol{X}(i)\boldsymbol{b}(i)]/(n-p-1)。 \tag{2-7}$$

ここで，回帰係数の最小二乗推定量にデータ $\{(y_i,\ \boldsymbol{x}_i)\}_{i=1}^{n}$ が与える影響を調べるため，第 $i$ 観測値にウェイトを付けたときの最小二乗推定量 $\boldsymbol{b}(w_i)$ を考察する。

$$\boldsymbol{b}(w_i)=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \tag{2-8}$$

$$\boldsymbol{W}=\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & w_i & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$

と与えられ，これを $w_i$ について微分すると，

$$\frac{\partial \boldsymbol{b}(w_i)}{\partial w_i}=\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'(y_i-\boldsymbol{x}_i\boldsymbol{b})}{[1-(1-w_i)\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i']^2} \tag{2-9}$$

さらに，$w_i=1$ とすると，

$$\left.\frac{\partial \boldsymbol{b}(w_i)}{\partial w_i}\right|_{w_i=1}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'(y_i-\boldsymbol{x}_i\boldsymbol{b}) \tag{2-10}$$

を得る。この式により $(y_i,\ \boldsymbol{x}_i)$ が回帰係数の最小二乗推定量に与える影響を調べることができる[2]。(2-10) の両辺の左側に $\boldsymbol{x}_i$ を乗ずると，

$$\boldsymbol{x}_i \cdot \frac{\partial \boldsymbol{b}(w_i)}{\partial w_i}\bigg|_{w_i=1} = \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'(y_i - \boldsymbol{x}_i b) \qquad (2\text{-}11)$$

のような形になる。

$\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'$ は行列 $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ の第 $i$ 対角要素であり，独立変数データの情報を与え，残差 $e_i = y_1 - \boldsymbol{x}_i\boldsymbol{b}$ は従属変数データの情報を与える。$\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'$ の値が大であるときの独立変数観測値は high leverage, $e_i$ が大であるときの従属変数観測値は異常値と呼ばれる③。

いま，$L(u_i|\boldsymbol{x}_i) = N(0,\ \sigma^2)$ を仮定すると，データ $\{(y_i,\ \boldsymbol{x}_i)\}_{i=1}^{n}$ は多変量正規分布 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I})$ から独立に抽出されたものである。このときの回帰係数の最小二乗推定量の効率は大である。しかし，観測値に異常値や high leverage が存在すると，その効率は低下する④。現実的には，推定値 $\boldsymbol{b}$ が歪んだ値（$\boldsymbol{\beta}$ からかけ離れた値）となってしまう。したがって，4節に記述する所定の方法に基づいて異常値や high leverage 独立変数観測値を検出し，それ等を除いてから回帰係数の最小二乗推定を行うことが考えられる。

**注**

② (2—10) は Belsley et *al.*〔**3**〕により提示された回帰係数の最小二乗推定量の influence function である。

③ 母集団分布とは異なる別の母集団から生じた観測値を異常値と名づける。

④ 効率の低下についての詳細な説明は，Fisher 情報行列を用いて行うことができる。

## 3 Hat 行列とその性質

異常値や独立変数の high leverage が回帰係数の最小二乗推定値を歪めたものとする。本節では，独立変数が従属変数の予測値および残差に影響することを示し，独立変数による行列 $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ の対角要素の性質

線型回帰モデルにおける異常値の検出と頑健推定におよぼす独立変数観測値の影響を検討する。

まず，従属変数の予測値は $\hat{\boldsymbol{y}}=\boldsymbol{X}\boldsymbol{b}$ で与えられるから，

$$\hat{\boldsymbol{y}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3-1}$$

ここで，$\hat{\boldsymbol{y}}$ の要素 $\hat{y}_i$ が一次関数となることから，(3-1) を

$$\begin{aligned}\hat{\boldsymbol{y}}&=\boldsymbol{H}\boldsymbol{y},\\ \boldsymbol{H}&=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',\end{aligned} \tag{3-2}$$

のように表わす。行列 $\boldsymbol{H}(n\times n)$ は $\boldsymbol{y}$ により $\hat{\boldsymbol{y}}$ の値を定める関数（写像）の作用をなし，これを “Hat 行列” と称す。

幾何学的にみると，ベクトル $\boldsymbol{y}$ と $\boldsymbol{X}$ の列を $n$ 次元ユークリット空間の点で表わすなら，点 $\boldsymbol{X}\boldsymbol{\beta}$（列ベクトル $\boldsymbol{x}_j$ の一次結合 $\sum_{j=1}^{p}\beta_j\boldsymbol{x}_j$ である）は $p$ 次元部分空間をなす。ベクトル $\hat{\boldsymbol{y}}$ は $\boldsymbol{y}$ に最も近いその部分空間の点であり，それはまた $\boldsymbol{y}$ のその部分空間への直交射影（perpendicular projection）である。したがって $\boldsymbol{H}$ は射影行列（projection matrix）である（Hoaglin and Welsch 〔12〕）。

また，行列 $\boldsymbol{H}$ は残差ベクトルにも現われて，

$$\begin{aligned}\boldsymbol{e}&=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\\ &=[\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{y},\end{aligned} \tag{3-3}$$

となる。独立変数から成る行列 $\boldsymbol{H}$ およびその要素 $h_{ij}$ のもつ性質は，Belsley et *al.*〔3〕と Hoaglin and Welsch 〔12〕の導出した結果に若干の補追を加えると以下のようである。

まず，$\hat{\boldsymbol{y}}$ の要素 $\hat{y}_i$ は，

$$\hat{y}_i=\boldsymbol{x}_i\boldsymbol{b}=\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$=h_{ii}\ y_i+\sum_{j\neq i}h_{ij}\ y_j, \qquad (3\text{-}4)$$
$$h_{ii}\equiv\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i',\ (\text{行列}\ \boldsymbol{H}\ \text{の第}\ i\ \text{対角要素})$$
$$h_{ij}\equiv\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_j',\ (\text{行列}\ \boldsymbol{H}\ \text{の}\ ij\ \text{要素})$$

のように表わされる。(3-4) は，行列 $\boldsymbol{H}$ の要素 $h_{ij}$ を梃子に $y_j$ が $\hat{y}_i$ に影響を及ぼすと解される。$\hat{y}_i$ に及ぼす $y_i$ の直接的な影響は対角要素 $h_{ii}$ を梃子にして現われる。

独立変数の観測値を採取していくと，行列 $\boldsymbol{X}$ のどの行ベクトルが，$\hat{\boldsymbol{y}}$ の該当する要素の値を大きくしているかを判定できる。行列 $\boldsymbol{H}$ が，その "high leverage point, $i$" を識別するために採用される。特に，その対角要素 $h_{ii}$ がそのためには有効である。次に $h_{ii}$ の性質を検討する。

行列 $\boldsymbol{H}$ は対称かつ巾等 (symmetric and idempotent) であるから ($\boldsymbol{H}^2=\boldsymbol{H}$)，対角要素は

$$h_{ii}=\sum_{j=1}^{n}h^2_{ij}=h^2_{ii}+\sum_{j\neq i}h^2_{ij}, \qquad (3\text{-}5)$$

となり，また $0\leq h_{ii}\leq 1$ である。巾等行列の固有値 (eigenvalues) は 0 か 1 のいずれかであり，また零ではない固有値の数はその行列の階数(rank)に等しいから，行列 $\boldsymbol{H}$ の場合には，$\text{rank}(\boldsymbol{H})=\text{rank}(\boldsymbol{X})=p$，である。巾等行列の rank と trace は等しいから，$\text{trace}(\boldsymbol{H})=p$，すなわち，

$$\sum_{i=1}^{n}h_{ii}=p \qquad (3\text{-}6)$$

なお，行列 $\boldsymbol{H}$ の対角要素の大きさの平均は，$p/n$，となる。

ところで，$l=[1\cdots 1]'$，$ll'=\boldsymbol{I}$，$\bar{\boldsymbol{y}}=\frac{1}{n}l'\boldsymbol{y}$，として $(\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}})$ を展開す

ると，

$$\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}=\boldsymbol{H}\boldsymbol{y}-l\bar{\boldsymbol{y}}$$
$$=\left[\boldsymbol{H}-\frac{1}{n}\boldsymbol{I}\right]\boldsymbol{y}=\tilde{\boldsymbol{H}}\boldsymbol{y} \tag{3-7}$$
$$\tilde{\boldsymbol{H}}=\boldsymbol{H}-\frac{1}{n}\boldsymbol{I} \tag{3-8}$$

となる。行列$\tilde{\boldsymbol{H}}$の要素 $\tilde{h}_{ij}$ は (3-8) から，

$$\tilde{h}_{ij}=h_{ij}-\frac{1}{n} \tag{3-9}$$

のようになる。ここで行列 $\boldsymbol{X}$ の各列 $\boldsymbol{x}_j=(x_{ij}\cdots x_{nj})'$ について平均をとり，$\bar{x}_j=\frac{1}{n}l'\boldsymbol{x}_j$，ベクトル$\bar{\boldsymbol{x}}=(\bar{x}_1\bar{x}_2\cdots\bar{x}_p)$ をつくる。そして，

$$\tilde{\boldsymbol{X}}=\begin{pmatrix}\boldsymbol{x}_1-\bar{\boldsymbol{x}}\\ \boldsymbol{x}_2-\bar{\boldsymbol{x}}\\ \vdots\\ \boldsymbol{x}_n-\bar{\boldsymbol{x}}\end{pmatrix}=\begin{pmatrix}x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2\cdots x_{1p}-\bar{x}_p\\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2\cdots x_{2p}-\bar{x}_p\\ \vdots & \vdots\\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2\cdots x_{np}-\bar{x}_p\end{pmatrix}=\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\\ \vdots\\ \tilde{\boldsymbol{x}}_n\end{pmatrix} \tag{3-10}$$

$$\tilde{\boldsymbol{H}}=\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}', \tag{3-11}$$

を定義する。行列 $\tilde{\boldsymbol{H}}$ の対角要素 $\tilde{h}_{ii}$ は，

$$\tilde{h}_{ii}=(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})' \tag{3-12}$$

であるから，行列 $\boldsymbol{H}$ の対角要素は (3-9) から

$$h_{ii}=\frac{1}{n}+(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})' \qquad (3\text{-}13)$$

のように与えられる。

$h_{ii}$ は $(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$ が大であるとき，即ち独立変数観測値の中でその平均値よりもはるかに大きな値を取るとき，大きくなる（1に近づく）。また $(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$ が小であるとき，小さな値となる（零に近づく）。

## 4 異常値の検出と独立変数観測値

線型回帰モデルでは誤差項を確率変数と仮定するので，(2-1) により従属変数は確率変数となる（独立変数は確率変数ではない）。従属変数の観測値が，（誤差項にたいして）仮定した確率分布と異なる別の分布の母集団から抽出された観測値（異常値）か否かの検定を行うとき，誤差項に正規性の仮定を置けば，Studentized residuals を用いることができる。この Studentized residuals は，

$$e_{si}=\frac{e_i}{\mathrm{S}(1-h_{ii})^{\frac{1}{2}}} \qquad (4\text{-}1)$$

と与えられるが，分子の残差 $e_i$ にも行列 $\boldsymbol{H}$ の対角要素 $h_{ii}$ が存在し，異常値の検出に影響を及ぼしている。

残差は (2-3) と (3-4) を用いて，

$$\begin{aligned} e_i&=y_i-\hat{y}_i \\ &=(1-h_{ii})y_i-\sum_{j\neq i}h_{ij}y_j, \end{aligned} \qquad (4\text{-}2)$$

と表わすことができる。しかし，異常値 $y_i$ は常に該当する残差 $e_i$ を大とするとはかぎらない。$h_{ii}$ が零に近い小さな値であるとき $e_i$ は大とな

るが，$h_{ii}$ が1に近い大きな値であるときには $e_i$ は小さな値となる。また，異常値 $y_i$ は該当する残差 $e_i$ ではなく，他の残差を大とする可能性がある。すなわち，もし $h_{mi}$ の値が大きければ，$y_i$ により残差 $y_m-\hat{y}_m$ の値が大きくなるからである。

残差 $e_i$ は $h_{ii}$ の減少関数となることから，Studentized residuals $e_{si}$ による異常値検定の検出力は，$h_{ii}$ の減少関数である[6]。$h_{ii}$ が大である場合の異常値を検出する能力は，$h_{ii}$ が小である場合に比べて劣ることになる。

ところで，線型回帰モデルの回帰係数の最小二乗推定量の漸近特性（大標本特性）については，Anderson〔**1**, pp.23-27〕と Eickel〔**8**〕によって取り扱かわれ，その漸近的正規性が証明されているが，Huber〔**15**〕は線型回帰係数の最小二乗推定量が漸近的正規性を保持するための必要十分条件は

$$\lim_{n\to\infty}(\max_i h_{ii})=0,\quad i=1,\ 2,\cdots,\ n \tag{4-3}$$

であることを示している。$h_{ii}$ の値が小さいことは，線型回帰係数の最小二乗推定量の漸近的正規性のためにも必要なことである。さらに Huber は (4-3) の条件はまた‘一種の“design（行列 $\boldsymbol{X}$）の頑健性”を保証する’とし，また‘この design の頑健性はすべての $i$ について $h_{ii}=p/n$ とすることによって最大化され，そのときに，design 行列 $\boldsymbol{X}$ が均斉している’としている（Huber〔**15**, p.804〕）。また，Box and Draper〔**5**〕は異常値に対して design を鋭敏でなくするためには，$h_{ii}$ を小さくすることだとしている。これらのことは，(3-4) の関係から，異常値が存在するとき $h_{ii}$ が大であると，これが梃子の作用をして予測値 $\hat{y}_i$ を大きくしてしまうことを警戒したものである。

(3-13) から明らかなように，$h_{ii}$ が大であることは，独立変数の第 $i$ 観測値 $(x_{i1},\ x_{i2}\cdots x_{ip})$ が大きい（high leverage）ことを示している。独

立変数の high leverage 観測値は Studentized residuals による異常値検出の能力を損い，また従属変数の予測値を肥大化する。経済変数のデータは時系列で与えられることが一般的であり，繰り返し測定したりコントロールすることが不可能であるから，行列$\boldsymbol{X}$は固定されたものとなる。したがって $h_{ii}$ を計算して，その値が大きい場合には該当する観測値 ($x_{i1}\cdots x_{ip}$) を除外しておくことが，上記の事態を回避するために適切な措置といえよう。

そのためには，$h_{ii}$ がどの程度の大きさとなったときに独立変数の観測値を削除するかを判定するための基準を設定する必要がある。Belsley et *al*.〔**3**〕は次のような方法で $h_{ii}>2p/n$ の場合に $i$ 番目の独立変数データを high leverage だと定めている。

独立変数 $\boldsymbol{x}_i$ が $p$ 次元の相互独立な正規分布にしたがうと仮定すると，$h_{ii}$ について正確な分布を計算することができる。独立変数についてそうした仮定を置くことは，現実的妥当性はないが，上記の目的のみに限定して導出した結果を用いる。その仮定の下で $h_{ii}$ について，

$$F=\frac{(n-p)}{(p-1)}\frac{\left\{h_{ii}-\left(\frac{1}{n}\right)\right\}}{(1-h_{ii})} \tag{4-4}$$

の関係は自由度 $p-1$ および $n-p$ の $F$ 分布にしたがうことが示される[6]。$p>10$ および $n-p>50$ について，$F$値の95%ポイントが2以下であるから，$2p/n$ ($h_{ii}$ の平均値の2倍) を $\boldsymbol{x}_i$ のカット・オフ基準とする。$p$が多く $n$ が少なく，$p/n>0.4$ の場合には基準が高くなり過ぎるので，すべての独立変数観測値に疑念が生じ，$p$が少ない場合にはカット・オフの対象となる独立変数観測値が少くなってしまう。しかし，$h_{ii}$ が $2p/n$ を越えた場合，$i$ 番目の独立変数観測値の位置を leverage point と称す。

注

⑤ Cook〔6〕は mean shift model において，第 $i$ 観測値が異常値ではないという仮説を検討するために，Gentleman and Wilk〔9〕の導出した統計量を基にF-統計量を算出している。この統計量は Studentized residuals の単調増加関数となっている。よって，この F-統計量による検定の検出力も $h_{ii}$ の減少関数である。

⑥ 詳細は付録を参照。

## 5 回帰係数の頑健推定と独立変数観測値

前節の方法のように，一定規準を越える異常値を除外して回帰係数を最小二乗推定するのでなく，頑健推定（robust estimation）と称する方法が開発されている。

現実問題への統計分析の経験を通して，観測値の母集団分布に正規性を仮定することに疑問が生じる場合がある。その中から生み出された頑健推定は，観測値の母集団を正規分布と仮定せず，また残差平方和最小化の規準を採らない推定方法である。その主たる狙いは，非常に大きな観測誤差に対して安全装置を設け，異常値が推定値に及ぼす影響を一定範囲におさめ，大きな異常値を隔離し，なおかつモデル（本稿の場合は線型回帰モデル）を最適なものとすることである。モデルに対する観測値の影響が，ほとんど検出不可能なものから，甚だしく大きなものまであることを考え，大半の観測値の動きに着目する。観測値母集団に特定の分布を仮定せず，また Gauss-Markov 定理を用いて分布の平均値を推定することをもって最適化を図るものでもない。

Huber〔15〕は回帰係数の頑健推定の方式を次のように示している。残差平方和の最小化，

$$\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i\boldsymbol{\beta})^2=\min. \qquad (5\text{-}1)$$

に代って，残差についての特定の関数を最小

$$\sum_{i=1}^{n}\rho(y_i-\boldsymbol{x}_i\boldsymbol{\beta})=\min \tag{5-2}$$

とするような $\boldsymbol{\beta}$ を推定する。ここで，$\rho(t)$ は凸関数 (convex function) であり Huber は，

$$\begin{aligned}\rho(t)&=\frac{1}{2}t^2 && |t|<c\\ &=c|t|-\frac{1}{2}c^2 && |t|\geq c\end{aligned} \tag{5-3}$$

を提示している。$c$ の値は許容しようと定めた $y_i$ の大きさに依って設定される。推定値は $c$ の値に対して敏感に反応しないので，誤差の標準偏差の1から2倍に $c$ を定めるのが合理的だとしている (Huber〔**16**〕)。

関数 $\rho(t)$ が微分可能であれば，$\Psi(t)=\rho'(t)$，(5-2) を満足するための必要条件は，

$$\sum_{i=1}^{n}x_{ij}\Psi(y_i-\boldsymbol{x}_i\boldsymbol{\beta})=0,\ \ j=1,\ 2,\cdots,\ P \tag{5-4}$$

である。

一般に，(5-4) は非線型方程式となるから，解を得るためには反復計算を必要とする。反復計算による推定値の求め方として，Newton 法，Bickel〔**4**〕による方法，Beaton and Tukey〔**2**〕による re-weighted least squares がある。いずれの方法を採るにしろ，関数 $\rho(t)$ では極小値に収束することのみが保証されるだけであるから，特性の良い初期値の選定が必要となる。

Holland and Welsch〔**13**〕は最小絶対値残差推定による初期値を用いて

re-weighted least squares を行うことを勧めている。最小絶対値残差推定量は漸近的正規性を有すので，反復計算により求めた回帰係数の頑健推定量もこの漸近特性を保持する。

ところで，Hampel〔10〕は各観測値 $\{(y_i,\ \boldsymbol{x}_i)\}$ が頑健推定に与える影響を示す influence function を定義しているが，Krasker and Welsch〔18〕にしたがってそれを次のように表わす。

回帰係数 $\boldsymbol{\beta}$ とその推定量 $\boldsymbol{b}$ について $\boldsymbol{b}-\boldsymbol{\beta}$ を考える⑦。そして，

$$\lim_{n\to\infty} Pr\left[\sqrt{n}\left\{(\boldsymbol{b}-\boldsymbol{\beta})-\frac{1}{n}\sum_{i=1}^{n}\Omega(y_i,\ \boldsymbol{x}_i)\right\}\right]=0 \qquad (5\text{-}5)$$

を満足する関数 $\Omega(y_i,\ \boldsymbol{x}_i)$ は推定量 $\boldsymbol{b}$ に対して $i$ 番目の観測値が与える影響を示す influence function である。もし，$E[\Omega(\boldsymbol{y},\ \boldsymbol{X})]=0$，$E|\Omega(\boldsymbol{y},\ \boldsymbol{X})|^2<\infty$ ならば，(4-5) により推定量は一致漸近的正規性を有し，その漸近分散は

$$V=E[\Omega(\boldsymbol{y},\ \boldsymbol{X})\Omega(\boldsymbol{y},\ \boldsymbol{X})'] \qquad (5\text{-}6)$$

となる。

Huber〔15〕の提示したタイプの回帰係数の頑健推定量の influence function は

$$\Omega(y_i,\ \boldsymbol{x}_i)=\Psi[y_i-\boldsymbol{x}_i\boldsymbol{\beta}](\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i' \qquad (5\text{-}7)$$

のようになる。$\Psi(o)$ が大きな値の残差を抑えても $\boldsymbol{x}_i'$ が乗ぜられているから，観測値の影響は任意に大となる。回帰係数の頑健推定では，独立変数の観測値中に他よりも特に大きな値のものがあるときには注意を要す。この点については Huber〔15〕が頑健推定にあたって，独立変数行列に課せられる条件として，

$$\max_{1\leqq i\leqq n} h_{ii} = \ll 1 \qquad (5\text{-}8)$$

を挙げている。

関数 $\rho(t)$ については幾つかのものが提示されており，モンテカルロ実験により各々の関数による頑健推定量の漸近的効率の比較的検討がなされている (Holland and Welsch [**12**])。どのような関数 $\rho(t)$ をもってしても，回帰係数の頑健推定の influence function は (5-7) の形になるから，独立変数の high leverage の影響は除去できない。

現実的な対応策として，$h_{ii}$ $(i=1\cdots n)$ を計算し，$2p/n$ を越える high leverage 独立変数観測値を除いて頑健推定を行うことが考えられる。

**注**

⑦ $\boldsymbol{b}$ が最小二乗推定量であり，また誤差項に正規性 $\boldsymbol{u}\sim N(0,\ \sigma^2\boldsymbol{I})$ の仮定を置くと，

$\boldsymbol{b}\sim N\ [\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$,

$\boldsymbol{b}-\boldsymbol{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$,

$E(\boldsymbol{b}-\boldsymbol{\beta})=0$,

$\mathrm{Var}(\boldsymbol{b}-\boldsymbol{\beta})=E[(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})']=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

## 5 むすび

線型回帰モデルの回帰係数の最小二乗推定量は一定条件の下では効率の大なる推定量である。しかし，観測値の母集団分布が裾野の厚いもの（正規分布ではない）であったり，観測値に異常値が存在するとき，また両者が合併したときには，最小二乗推定量の効率は著しく低下する。

観測値母集団に正規性が仮定できる場合，Studentized residuals を用い異常値を検出，除去した後に最小二乗推定を行う。正規性の仮定が設定できず，かつ異常値の影響を排除しようとする場合，頑健推定の方法が開発されている。いずれの方法においても，high leverage 独立変数観測値

の影響を除去できないので，$h_{ii}$ の計算によって high leverage point を検出し，該当する独立変数観測値を除外することが現実的対応であろう。

回帰係数の推定にあたって，異常値や high leverage 独立変数の影響を除去するため，極く一部の観測値が及ぼす影響を制限しながら回帰係数の推定を行い，その推定量が漸近的正規性を保有しかつその漸近分散をできるだけ小さいものとする試みがなされている (Krasker and Welsch [**18**])。経済変数間の関係を回帰分析する場合，この推定方法を採用する問題については次の課題としたい。

## 付　　録

### 1　$\tilde{\boldsymbol{x}}_i$ の分布

いま $\boldsymbol{X}$ を多次元正規分布にしたがう確率変数と仮定する。そして次のように $\tilde{\boldsymbol{X}}$ を定義する。

$$\tilde{\boldsymbol{X}}=\begin{pmatrix}\boldsymbol{x}_1-\bar{\boldsymbol{x}}\\ \boldsymbol{x}_2-\bar{\boldsymbol{x}}\\ \vdots\\ \boldsymbol{x}_i-\bar{\boldsymbol{x}}\\ \vdots\\ \boldsymbol{x}_n-\bar{\boldsymbol{x}}\end{pmatrix}=\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\\ \vdots\\ \tilde{\boldsymbol{x}}_i\\ \vdots\\ \tilde{\boldsymbol{x}}_n\end{pmatrix}=\begin{pmatrix}x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2\cdots x_{1p}-\bar{x}_p\\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2\cdots x_{2p}-\bar{x}_p\\ \vdots & \vdots\\ x_{i1}-\bar{x}_1 & x_{i2}-\bar{x}_2 \quad x_{ip}-\bar{x}_p\\ \vdots & \vdots\\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 \quad x_{np}-\bar{x}_p\end{pmatrix}$$

$$\bar{\boldsymbol{x}}=(\bar{x}_1\bar{x}_2\cdots\bar{x}_p),$$

$$\bar{x}_j=\frac{1}{n}l'\boldsymbol{x}_j,\quad j=1\cdots p,\quad l=(1\cdots1)'(n\times1),$$

$$\boldsymbol{x}_j=(x_{1j}x_{2j}\cdots x_{nj})',$$

$$\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}=\sum_{i=1}^{n}\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{x}}_i,$$

$$\bar{\tilde{x}}=\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i=0,\quad(\because\quad\sum\tilde{x}_i=0)$$

$$\bar{\tilde{x}}(i)=\frac{1}{n-1}\sum_{k\neq i}x_k=(n\bar{\tilde{x}}-\tilde{x}_i)/(n-1)=(\bar{\tilde{x}}-\tilde{x}_i/n)/(n-1)/n$$

ここで，$\tilde{x}_i$ の影響を調べるために，$\tilde{X}$ から $\tilde{x}_i$ を除いた $\tilde{X}(i)$ の分散は，

$$V\{\tilde{X}(i)\}=\frac{1}{n-1}\sum_{k\neq i}x_k'x_k-\bar{\tilde{x}}'(i)\bar{\tilde{x}}(i)=\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{x}_i'\tilde{x}_i-\tilde{x}_i'\tilde{x}_i\right)$$
$$-\bar{\tilde{x}}'(i)\bar{\tilde{x}}(i)$$

また，$\tilde{X}$ の分散は $\bar{\tilde{X}}=0$ であるから

$$V(\tilde{X})=\frac{1}{n-P}\sum_{i=1}^{n}\tilde{x}_i'\tilde{x}_i$$

である。$V\{\tilde{X}(i)\}$ と $V(\tilde{X})$ についての Wilk's $\Lambda$ statistics (Rao [**19**, p. 570]) は，

$$\Lambda(\tilde{x}_i)=\frac{\det\{\tilde{X}'\tilde{X}-(n-1)\bar{\tilde{x}}'(i)\bar{\tilde{x}}(i)-\tilde{x}_i'\tilde{x}_i\}}{\det(\tilde{X}'\tilde{X})}$$

となる。分子は

$$\det\left\{\tilde{X}'\tilde{X}-\frac{n^2}{(n-1)}\left(\bar{\tilde{x}}-\frac{\tilde{x}_i}{n}\right)'\left(\bar{\tilde{x}}-\frac{\tilde{x}_i}{n}\right)-\tilde{x}_i'\tilde{x}_i\right\}$$

となり，$\bar{\tilde{x}}=0$ であるから

$$\det\left\{\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}-\frac{n}{n-1}\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{x}}_i\right\}$$

さらに，

$$\det\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}-\frac{n}{n-1}\tilde{\boldsymbol{x}}_i'\tilde{\boldsymbol{x}}_i\right)=\left(1-\frac{n}{n-1}\tilde{\boldsymbol{x}}_i'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{x}}_i\right)\det(\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}})$$

$$=\left(1-\frac{n}{n-1}\tilde{h}_{ii}\right)\det(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}),$$

よって $\Lambda(\tilde{\boldsymbol{x}}_i)$ は，

$$\Lambda(\tilde{\boldsymbol{x}}_i)=1-\frac{n}{n-1}\tilde{h}_{ii}=\frac{n}{n-1}(1-h_{ii})$$

となる。

$\Lambda(\tilde{\boldsymbol{x}}_i)$ は $\chi^2$ 分布にしたがう統計量であり，この値が大ならば（小ならば）$\tilde{\boldsymbol{x}}$ の影響が小（$\tilde{\boldsymbol{x}}$ の影響大）となる。$\tilde{\boldsymbol{X}}$ は $p$ 次元正規分布からの $n$ 個の独立な標本で成るから，$\Lambda(\tilde{\boldsymbol{x}})$ は $F$ 分布にとして次のように表わせる(Rao〔**19**, p.570〕)。

$$\frac{n-p}{p-1}\left[\frac{1-\Lambda(\tilde{\boldsymbol{x}}_i)}{V(\tilde{\boldsymbol{x}}_i)}\right]\sim F_{p-1,\ n-p}$$

したがって，

$$\frac{n-p}{p-1}\ \frac{[h_{ii}-(1/n)]}{(1-h_{ii})}\sim F_{p-1,\ n-p}.$$

（1986年3月）

## REFERECES

〔1〕 Anderson, T. W. : *The Statistical Analysis of Time Series*. New York : John Wiley & Sons, 1971.

〔2〕 Beaton, A. E. and J. W. Tukey. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16 (1974), 147-185.

〔3〕 Belsley, D. A., E. Kuh and R. E. Welsch : *Regression Diagn ostics ; Identifying Influential Data and Sources of Collinearity*. New York : John Wiley & Sons, 1980.

〔4〕 Bickel, P. J. : "One Step Huber Estimates in the Linear Model," *Journal of the Amercan Statistical Association*, 70 (1975), 428-434.

〔5〕 Box, G. E. P. and N. R. Draper : "Robust Designs", *Biometrika*, 62(1975), 347-351.

〔6〕 Cook, R. D. : "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74 (1979), 169-174.

〔7〕 Cook, R. D. and S. Weisberg : "Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22 (1980), 495-508.

〔8〕 Eicker, F. : "Asymptotic Normarity and Consistency of the Least Squares Estimators for Families of Linear Regressions", *The Annals of Mathematical Statistics*, 34 (163), 447-456.

〔9〕 Gentleman, J. F. and M. B. Wilk : "Detecting Outliers Ⅱ. Supplementing the Direct Analysis of Residuals", *Biometrics*, 31 (1975), 387-410.

〔10〕 Hampel, F. R. : "Robust Estimation : A Condensed Partial Survey", *Zeitschrift für Wahrscheinlichkeitstheorie und Verwarndte Gebiete*, 27 (1973), 87-104.

〔11〕 ——— : "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69 (1974), 383-394.

〔12〕 Hoaglin, D. C. and R. E. Welsch : "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32(1978), 17-22.

〔13〕 Holland, P. W. and R. E. Welsch : "Robust Regression Using Iteratively Reweighted Least Squares," *Communications in Statistics*, A6 (1977), 813-827.

〔14〕 Huber, P. J. : "Robust Statistics : A Review," *The Annals of Mathematical Statistics*, 43 (1972), 1041-1067.

〔15〕 ——— : "Robust Regression : Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1 (1973), 799-821.

〔16〕 ——— : "Robustness and Designs," in *A Survey of Statistical Design and Linear Models* : J. N. Stristava, ed., Amsterdam : North Holland, 1975, 287-301.

〔17〕 Krasker, W. S. : "Estimation in Linear Regression Models with Disparate Data Points," *Econometrica*, 48 (1980), 1333-1346.

〔18〕 Krasker, W. S. and R. E. Welsch : "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 77 (1982), 595-604.

〔19〕 Rao, C. R. : *Linear Statistical Inference and Its Applications*, 2nd ed.. New York : John Wiley & Sons, 1973.

〔20〕 Theil, H. : *Principles in Econometrics*. Amesterdam : North Holland, 1979.